# AI Accountability Policy Request for Comment National Telecommunications and Information Administration, U.S. Department of Commerce

Julia Stoyanovich, New York University, stoyanovich@nyu.edu

## Respondent information

The Center for Responsible AI at New York University, *NYU R/AI*, was established in Fall 2020 under the leadership of Prof. Julia Stoyanovich, Institute Associate Professor of Computer Science and Engineering at the NYU Tandon School of Engineering and Associate Professor of Data Science at the NYU Center for Data Science.

*NYU R/AI* is a comprehensive laboratory for accelerating responsible AI practices that is building a future in which "responsible AI" is synonymous with "AI". Its mission is to advance broad adoption of responsible AI through rigorous multi-disciplinary research, cross-sector dialogue, and technical collaborations to create an ecosystem of equitable AI design, development, deployment, and oversight by data scientists, decision-makers, and the public. *NYU R/AI*'s interdisciplinary team includes academic researchers, educators, and practitioners with expertise in computer science, data science, sociology, and technology policy. This response for comment is based on the research and practical experience of the *NYU R/AI* team, summarized below.

- Extensive peer-reviewed technical and socio-technical responsible AI research of the *NYU R/AI* team [45, 32, 20, 18, 5, 28, 30, 8, 41, 3, 47, 48, 9, 17, 7, 46, 52, 27, 14, 24, 42, 51, 2, 22, 43, 19, 13, 26, 25, 53, 54, 49, 16, 31, 15, 50, 34, 12, 21, 10].
- Stoyanovich's collaboration with the Office of the New York State Comptroller on an AI governance audit. The objective of this audit was to assess New York City's progress in establishing an appropriate governance structure over the development and use of artificial intelligence (AI) tools and systems. The audit covered the period from January 2019 through November 2022. The key finding of the audit is that "NYC does not have an effective AI governance framework." [1]
- Stoyanovich's advocacy for the New York City Local Law 144 of 2021 "in relation to Automated Employment Decision Tools (AEDTs)". [2] [3] [4] [5] [6] [7]

[1] https://www.osc.state.ny.us/state-agencies/audits/2023/02/16/artificial-intelligence-governance

[2] https://www.nytimes.com/2021/03/17/opinion/ai-employment-bias-nyc.html

[3] https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

[4] https://www.nytimes.com/2023/05/25/technology/ai-hiring-law-new-york.html

[5] https://rules.cityofnewyork.us/wp-content/uploads/2022/12/Stoyanovich_144_Jan23_2023.pdf

[6] https://dataresponsibly.github.io/documents/Stoyanovich_Int1894Testimony.pdf

[7] https://dataresponsibly.github.io/documents/Bill1894Showreel.pdf

- Stoyanovich's work on the New York City Automated Decision Systems (ADS) task force, by Mayoral appointment.
- Public education work of the *NYU R/AI* team, including a course "We are AI: Taking control of technololgy," developed in collaboration wtih the Queens Public Library and P2PU, and offered regularly to members of the public [8] and two comic book series: "We are AI" (in English and in Spanish) [38, 36, 44, 35, 37] and "Data Responsibly" [6, 4].
- Education of data science students and practitioners, including responsible data science courses [9], and the algorithmic transparency course [10] and playbook. [11]

# Response to specific questions

While it is tempting to provide an opinion on each question, particularly because these questions are closely related, we decided to focus our attention on specific questions where the *NYU R/AI* team has experience and expertise to contribute.

## AI Accountability Objectives

**1. What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments?**

The purpose of AI accountability mechanisms is to create a distributed accountability regime in which different stakeholders share responsibility for the design, development, use, and oversight of AI. Certifications, external and internal audits, and assessments — of the over-all and case-specific performance of an AI-enabled system, of its benefits, of its potential risks and actual harms — all make part of this regime.

The most powerful and important stakeholder group is "society at large" —- the totality of individuals being impacted by AI systems in their everyday lives. It is crucial to involve this group in deliberations about appropriate " legal standards and enforceable risk thresholds" before these standards and thresholds are put in place, and to continue these deliberations throughout the use of these systems. For this reason, it is crucial to implement accountability mechanisms early and to revisit them often.

A mechanism that is both very powerful and very simple to enact is notifying people when they are interacting with / being subjected to AI-assisted decisions [11]. Such a mechanism must be in place whenever AI is used in decision-making.

[8] https://dataresponsibly.github.io/we-are-ai/

[9] https://dataresponsibly.github.io/courses/

[10] https://dataresponsibly.github.io/algorithmic-transparency-playbook/

[11] https://dataresponsibly.github.io/algorithmic-transparency-playbook/resources/transparency_playbook_camera_ready.pdf

**3. AI accountability measures have been proposed in connection with many different goals, including those listed below. To what extent are there tradeoffs among these goals? To what extent can these inquiries be conducted by a single team or instrument?**

As noted in response to #1, accountability has to be distributed. For this reason, no single accountability measure will suffice, and no single team or instrument will be sufficient.

**5. Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?**

As noted in response to #1, the first requirement is to notify people that they are interacting with an AI, rather than (or in addition to) a human. The second requirement is to explain what goals the AI pursues (e.g., to help you apply for jobs faster, or to help you find as many job openings as possible that may be a good fit but don't immediately match the qualifications stated in your resume). The third requirement is to explain how the AI is assessed against these goals (i.e., does it work and how do we know whether it works). These basic requirements are the same for simple AI systems like those used in scoring and ranking [39] and for more complex systems like those that use large language models. These steps are crucial, before more sophisticated types of assessment are conducted.

Conveniently, in the case of ChatGPT, one can ask what goals the system pursues. In Spring 2023, OpenAI's chatGPT would respond that its goals is to sound like a human, rather than to give correct answers to questions. This is important to know for people who are interacting with ChatGPT and expect to learn factual information.

**7. Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? [...]**

Unfortunately, the use of AI remains unaccountable even in critical domains like credit and lending, housing, and hiring and employment, and many others. The public remains largely uninformed about the extent to which AI is already embedded into decision-making. Because so little is currently disclosed about the use of AI, any amount of disclosure will be seen as positive. However, disclosing partial (and often explicitly cherry picked) information can be misleading.

As an example, NYC Local Law 144 of 2021 (LL 144) aims to strengthen accountability in the use of automated decision systems in hiring and employment. Critics of this law are rightfully concerned about the very limited scope of both auditing and disclosure that this law mandates. Passing a bias audit and meeting the disclosure requirements mandated by LL 144 will legitimize tools that are otherwise ineffective (e.g., lack validity) or discriminatory (e.g., with respect to age, which falls out of scope of the bias audit).

To counteract such effects, we must raise the overall expectations regarding the accountability of automated decision-making. A meaningful first step is to strengthen requirements for public disclosure about the use of AI, the goals that it pursues, and the assessment against these goals, per response #5 above.

Anecdotally, when the opinion piece titled "We need laws to take on racism and sexism in hiring technology" by Givens, Schellmann and Stoyanovich appeared in the New York

Times in March 2021, many readers commented that other demographic groups (e.g., those based on age) should be considered. Another noteworthy comment by one of the readers, referring to Pymetrics' games-based assessment from the article, was: "I would not blow up a balloon just to get hired. This test would tell me the employer is stupid. "

## Existing Resources and Models

**9. What AI accountability mechanisms are currently being used? Are the accountability frameworks of certain sectors, industries, or market participants especially mature as compared to others? Which industry, civil society, or governmental accountability instruments, guidelines, or policies are most appropriate for implementation and operationalization at scale in the United States? Who are the people currently doing AI accountability work?**

The request for comment already refers to several lines of work on accountability mechanisms, including datasheets for datasets and model cards. Below, we first discuss important work done by government accountability professionals in which we have been directly involved. Then, we discuss recent work by members of team a specific accountability frameworks and on their domain-specific instantiations.

**Office of the New York State Comptroller recently conducted an audit of AI governance in New York City's agencies** (*NYU R/AI*'s Stoyanovich served as a consultant) [12] The goal of the audit was to assess NYC's progress in establishing an appropriate governance structure over the development and use of AI tools and systems. The audit covered the period from January 2019 through November 2022, and focused its attention on a sample of four City agencies: NYC Police Department (NYPD), Administration for Children's Services (ACS), Department of Education (DOE), and Department of Buildings (DOB).

The key finding was that NYC does not have an effective AI governance framework. While agencies are required to report certain types of AI use on an annual basis, there are no rules or guidance on the actual use of AI. Consequently, City agencies developed their own, divergent approaches. The audit found ad hoc and incomplete approaches to AI governance, "which do not ensure that the City's use of AI is transparent, accurate, and unbiased and avoids disparate impacts."

Some agencies perform certain activities that partially address components of AI governance, such as identifying appropriate use, intended outcomes, data governance, and potential impacts, but do so because of laws created to address issues not specific to AI.

Sone specific findings are summarized here, and detailed in the report [13].

An important finding of the audit concerns bias in decision-making. ACS has taken specific steps to address possible bias in its Severe Harm Predictive Model by eliminating certain types of racial and ethnic data and testing the model's output against benchmarks. In contrast, DOE does not require any steps to determine whether the AI tools available to its schools have been evaluated to address potential bias.

[12] https://www.osc.state.ny.us/state-agencies/audits/2023/02/16/artificial-intelligence-governance

[13] https://www.osc.state.ny.us/files/state-agencies/audits/pdf/sga-2023-21n10.pdf

Another important finding concerns the use of facial recognition technology by the NYPD, with respect to both bias and effectiveness (accuracy). The NYPD created impact and use policies for its surveillance tools in compliance with the NYC Public Oversight of Surveillance Technology Act. The impact and use policy of its facial recognition technology acknowledges the potential bias of facial recognition, particularly against groups other than white males. It further states that NYPD only uses facial recognition technology that has been evaluated by the National Institute of Standards and Technology (NIST). However, NYPD did not review NIST's evaluation of the facial recognition technology it used, nor did it establish what level of accuracy would be acceptable. NYPD officials explained that any potential match is reviewed by multiple individuals to help mitigate potential accuracy and bias issues.

This audit also highlights the challenges that New York City has been facing in consistently following through on an effort to make AI-assisted government decision-making transparency and accountable. The first effort in this regard — in New York City as well as nationally — was the passing of Local Law 49 of 2018, which required the establishment of an "Automated Decision Systems (ADS) Task Force." [14] (*NYU R/AI*'s Stoyanovich was an appointed member of the task force.) The task force issued its recommendations in late 2019. [15] In response to the report, Mayor de Blasio issued an Executive Order (EO 50 of 2019) to establish an Algorithms Management and Policy Office (AMPO). AMPO's task was to create a reporting framework of algorithmic tools, policies, and protocols to guide the City and its agencies in the fair and responsible use of such tools, a process for individuals to learn about the City's use of these tools, a complaint resolution process for those impacted by such use, and a public education strategy. AMPO created an initial list of tools, but fell short of establishing comprehensive policies and protocols. In January 2022, AMPO was discontinued by Executive Order 3 of Mayor Adams.

**Nutritional labels for automated decision systems.** We have been developing accountability mechanisms that are based on the concept of a "nutritional label" [39, 40] for automated decisions. We have extended this mechanism to consider data, process, and outcomes, and to speak to the needs of different stakeholders [41, 45]. Nutritional labels can be used to support public-facing disclosure requirements, such as those of Local Law 144. Such standardized labels can list the factors that go into a tool's decision, both before candidates are screened and after a decision on their application is made. Job seekers, employees, and their representatives should be directly involved in the design and testing of such labels. Figure 3 gives an example of a possible

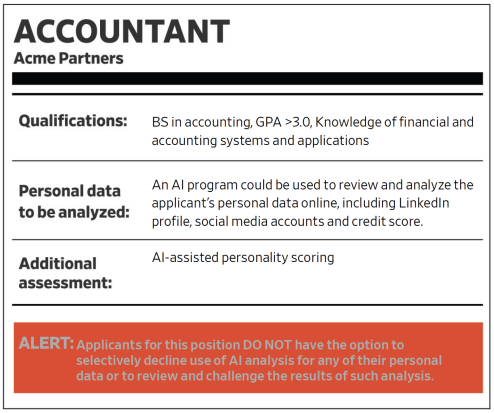


Figure 1: A "posting label" to accompany a job posting for an accountant role.

[14] https://www.nyc.gov/site/adstaskforce/index.page

[15] https://www.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf

"posting label" that would accompany a job ad and contain a short and clear summary of the screening process. [33]. This label is presented to a job seeker before they apply, supporting informed consent, allowing them to opt out of components of the process, or to request accommodations. Giving an opportunity to request accommodations is particularly important in light of the recent guidance by the Equal Employment Opportunity Commission on the Americans with Disabilities Act and the use of AI to assess job applicants and employees. [16]

**Contextual transparency for automated decision systems.** In a recent paper [32], *NYU R/AI*'s Sloane and Stoyanovich introduced the concept of contextual transparency as an approach that integrates social science, engineering and information design to improve ADS transparency for specific professions, business processes and stakeholder groups. We demonstrate the applicability of contextual transparency by using it to design a nutritional label for human resources professionals.

LinkedIn Recruiter is a real-world ADS example —– it selects candidates that fit the criteria specified by a recruiter —– but different professions use ADS in different ways. To make a transparency mechanism context-specific, we recommend using three contextual transparency principles (CTP):

- CTP 1: Social Science for Stakeholder Specificity: This aims to identify the professionals who rely on a particular ADS, how exactly they use it, and what information they need to have about the system to do their jobs better. This can be accomplished through surveys or interviews.
- CTP 2: Engineering for ADS Specificity: This aims to understand the technical context of the ADS used by the relevant stakeholders. Different types of ADS operate with different assumptions, mechanisms and technical constraints. This principle requires an understanding of both the input, the data being used in decision-making, and the output.
- CTP 3: Design for Transparency- and Outcome-Specificity: This aims to understand the link between process transparency and the specific outcomes the ADS would ideally deliver. In recruiting, for example, the outcome could be a more diverse pool of candidates facilitated by an explainable ranking model.

We also conducted empirical work, focusing on the profession of recruiting. Based on our findings, and on the contextual transparency principles discussed above, we created a prototype of a nutritional label for human resources professionals, specifically, for those who source and screen applicants with the help of LinkedIn Recruiter. This label would be inserted into the typical workflow of LinkedIn Recruiter users, allowing them to both assess the degree to which the ranked results satisfy the intent of their original search, and to refine their search to generate better results.

[16] https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence

**14. Which non-U.S. or U.S. (federal, state, or local) laws and regulations already requiring an AI audit, assessment, or other accountability mechanism are most useful and why? Which are least useful and why?**

We already discussed NYC's LL 144 and LL 49 (see response to #9). Both of these are important and useful examples that underscore, first and foremost, the importance of multi-stakeholder deliberation around accountability. In the case of LL 144, commercial vendors were by far over-represented in the conversation, and job seekers and the public at large were under-represented.

In her testimony during one of the rule-making hearings on this law, Stoyanovich argues for the need to involve job seekers, employees, and their representatives in defining standards for [Automated Employment Decision Tools] AEDT audits and notices: [17]

"Local Law 144 is an incredible opportunity for New York City to lead by example, but only if this law is enacted in a way that is responsive to the needs ofall key stakeholders. The conversation thus far has been dominated by the voices of commercial entities, especially by AEDT vendors and organizations that represent them, but also by employers who use AEDT, and by commercial entities wishing to conduct AEDT audits. However, as is evident from the fact that we are testifying in front of the Department of Consumer and Worker Protection, the main stakeholder group Local Law 144 aims to protect — from unlawful discrimination, and arbitrary and capricious decision-making — are job candidates and employees. And yet, their voices haven't been heard prominently in the conversation! New York City must ensure active participation of a diverse group of job seekers, employees and their representatives in both rule making and enactment of Local Law 144. [...]"

Another shortcoming of Local Law 144 in its current form is that its public disclosure component has been weakened during rule making. The original law required that features and characteristics based on which a tool is making its determination be disclosed. This helps get at the question of validity. However, in its current form, the law does not require such disclosure. It also does not contain any provisions to check whether the tool is effective (i.e., whether it works). Future accountability mechanisms should use Local Law 144 as an example, to see what works and what does not, and to also include stronger provisions for surfacing information about effectiveness of the tools, and reasons for the decisions made by the tools.

[17] Recommendations 1 and 3 in her testimony, see https://rules.cityofnewyork.us/wp-content/uploads/2022/12/Stoyanovich_144_Jan23_2023.pdf.

## Accountability Subjects

**16. The lifecycle of any given AI system or component also presents distinct junctures for assessment, audit, and other measures. For example, in the case of bias, it has been shown that "[b]ias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed—or if AI is required at all." How should AI accountability mechanisms consider the AI lifecycle?**

Question #16 and several questions in this and next sections pertain to accountability throughout the lifecycle, and we will tackle them jointly. To start, although there is no single formalization of "the lifecycle," we typically invoke the lifecycle framing when attempting to bridge the methodological divide between a reductionist and a holistic view of the AI system. How specifically a lifecycle is conceptualized (e.g., what the "boxes" and the "arrows" denote) depends on the point of view regarding which components can be decoupled from which other components.

Concretely, *NYU R/AI* has been engaged in technical work that pertains to a very specific lifecycle: the lifecycle of data collection, analysis, and use [43, 34]. The view is still reductionist, in that it assumes that a decision was made to collect specific kinds of data and build an ADS that uses that date. However, compared to the input/model/output view that is often taken in the technical fairness, accountability, and trasnaprency research, this view allows us to expand our focus beyond a single data analysis module (e.g., a predictor), and to consider the impacts of the technical choices made during data collection, curation, integration, and other types of pre-processing (e.g., missing value imputation) on the properties of predictors downstream from these choices.

For example, suppose that the feature "number of years of experience" is used (by an AI) to determine the level of compensation to offer to a job applicant. If the value of this feature is missing, it will have to be filled in ("interpolated") during data preprocessing. If the data scientist uses the default data interpolation method, they will replace the missing values by the median for the population. This, however, will guess values incorrectly for the more experienced (i.e., older) applicants — it will skew them "younger." The reason for this is that older people are more likely to worry about ageism, and will thus attempt to mask or omit any information that is a strong proxy for age. Consequently, whether a value of the "years of experience" feature is missing depends on what that value actually is (i.e., values are not "missing at random") — higher values will be missing more frequently. If this skey is introduced during data pre-processing and not detected, it will lead to consequences downstream, namely, to making offers with lower compensation to older applicants.

Understanding the impact of data pre-processing choices on model performance, as the preceeding example illustrates, is an active area of research. Some tools are being developed by us and others [15] that can help assess whether, for example, the proportion of members of demographic groups changes (often inadvertently) as data is transformed — for purely technical reasons — during preprocessing, or whether the accuracy of prediction for different demographic groups changes depending on how missing values are interpolated [16, 31, 15]. Our insights are that even the seemingly mundane and value-

neutral technical choices matter! Based on these choices, effectiveness, fairness, and robustness of predictions can change. For this reason, technical accountability primitives have to be integrated directly into the data lifecycle management.

**18. Should AI systems be released with quality assurance certifications, especially if they are higher risk?**

Yes, AI systems should be released with quality assurance certifications. AI systems are engineering artifacts. They were were not tested according to well-defined criteria, and if we have no way to check that the results of those tests are satisfactory, then we cannot take the claim that they work on faith.

There is strong evidence to suggest that recommendations of many of these tools are inconsistent and arbitrary. Let us consider hiring and employment. Tools that don't work hurt job seekers and employees, subjecting them to arbitrary decision-making with no recourse. Tools that don't work also hurt employers, they waste money paying for software that doesn't work, and miss out on many well-qualified candidates based on a self-fulfilling prophecy delivered by a tool.

In our own work, done in collaboration with an interdisciplinary team that included several data scientists, a sociologist, an industrial-organizational (I-O) psychologist, and an investigative journalist, we evaluated the validity of two algorithmic personality tests that are used for pre-employment assessment. We conducted an external audit of two tools — released by Humantic AI and Crystal — that claim to construct "personality profiles" of job seekers based on their resume, or LinkeIn profile, or Twitter handle [26, 25]. Importantly, rather than challenging or affirming the assumptions made in psychometric testing—that personality traits are meaningful and measurable constructs, and that they are indicative of future success on the job—we framed our methodology around testing the assumptions made by the vendors themselves. We found that both tools show substantial instability on key facets of measurement, and so cannot be considered valid testing instruments.

| Facet | Crystal | Humantic |
|---|---|---|
| Resume file format | ✗ | ✓ |
| LinkedIn URL in resume | ? | ✗ |
| Source context | ✗ | ✗ |
| Algorithm-time / immediate | ✓ | ✓ |
| Algorithm-time / 31 days | ✓ | ✗ |
| Participant-time / LinkedIn | ✗ | ✗ |
| Participant-time / Twitter | N/A | ✓ |

Figure 2: Summary of stability results for Crystaland Humantic AI, see [26] for details.

For example, Crystal frequently computes different personality profiles if the same resume is given in PDF vs. in raw text, while Humantic AI gives different personality profiles on a LinkedIn profile vs. a resume of the same job seeker, violating the assumption that the output of a personality test is stable across job-irrelevant input variations. Such tools cannot be allowed to proliferate, and accountability mechanisms should help protect candidates and employers from their use!

## Accountability Inputs and Transparency

**20. What sorts of records (e.g., logs, versions, model selection, data selection) and other documentation should developers and deployers of AI systems keep in order to support AI accountability? How long should this documentation be retained? Are there design principles (including technical design) for AI systems that would foster accountability-by-design?**

We discussed "nutritional labels" in the answer to #9, highlighting their usefulness as a public-facing accontability method. Nutritional labels can be used more generally to support accountaiblity requirements throughout the data lifecycle (discussed briefly in the answer to question #16), fostering accountability-by-design [41].

The technical data management and cyberinfrastructure communities have been studying systems and standards for metadata, provenance, and transparency for decades [1, 23]. We are now seeing renewed interest in these topics due to the proliferation of data science applications that use data opportunistically. Several recent projects explore these concepts for data and algorithmic transparency, including the Dataset Nutrition Label project, Datasheets for Datasets, Model Cards, System Cards, among others. All these method rely on manually constructed annotations. In contrast, our goal is to generate labels automatically or semi-automatically.

To differentiate a nutritional label from more general forms of metadata, we articulate several properties:

- **Comprehensible**: The label is not a complete (and therefore overwhelming) history of every processing step applied to produce the result. This approach has its place and has been extensively studied in the literature on scientific workflows, but is unsuitable for the applications we target. The information on a nutritional label must be short, simple, and clear.
- **Consultative**: Nutritional labels should provide actionable information, rather than just descriptive metadata. For example, universities may invest in research to improve their ranking, or consumers may cancel unused credit card accounts to improve their credit score.
- **Comparable**: Nutritional labels enable comparisons between related products, implying a standard.
- **Concrete**: The label must contain more than just general statements about the source of the data; such statements do not provide sufficient information to make technical decisions on whether or not to use the data.

Data and models are chained together into complex automated pipelines — computational systems "consume" datasets at least as often as people do, and therefore also require nutritional labels. We articulate additional properties in this context:

- **Computable**: Although primarily intended for human consumption, nutritional labels should be machine-readable to enable specific applications: data discovery, integration, automated warnings of potential misuse.

- **Composable**: Datasets are frequently integrated to construct training data; the nutritional labels must be similarly integratable. In some situations, the composed label is simple to construct: the union of sources. In other cases, the biases may interact in complex ways: a group may be sufficiently represented in each source dataset, but underrepresented in their join.
- **Concomitant**: The label should be carried with the dataset; systems should be designed to propagate labels through processing steps, modifying the label as appropriate, and supporting the paradigm of accountability-by-design.

**21. What are the obstacles to the flow of information necessary for AI accountability either within an organization or to outside examiners? What policies might ease researcher and other third-party access to inputs necessary to conduct AI audits or assessments?**

One important obstacle is that data on which AI systems are trained or tested, or data on which these systems operate in deployment, is sensitive for privacy reasons and cannot be released directly. Privacy-preserving synthetic data generation is powerful, and techniques from this domain are rapidly maturing. In our own recent work, we showed that differentially private synthetic data can achieve epistemic parity with real data in research. Specifically, we showed that it is often possible to reproduce results of peer-reviewed social science papers over privacy-preserving synthetic versions of that (public) datasets that were analyzed in these papers [29]. We developed an open-source epistemic parity benchmark [18] that will grow over time, to include additional datasets, data generators, and scientific papers to be reproduced.

**23. How should AI accountability "products" (e.g., audit results) be communicated to different stakeholders? Should there be standardized reporting within a sector and/or across sectors? How should the translational work of communicating AI accountability results to affected people and communities be done and supported?**

See response to questions #9 and #20.

## Barriers to Effective Accountability

**24. What are the most significant barriers to effective AI accountability in the private sector, including barriers to independent AI audits, whether cooperative or adversarial? What are the best strategies and interventions to overcome these barriers?**

One of the most significant barriers to effective AI accountability, in both the private and the public sectors, is the lack of responsible AI literacy. As part of ongoing work, we conducted an extensive review of AI literacy initiatives and found that most of them focus on teaching technical (rather than socio-technical) concepts, and are targeted to K-12 students. This leaves out those who are primarily interested in learning about the socio-legal impacts of AI (rather than learning how to code). And it leaves out technical practitioners who may be well-versed in the technical aspects of AI but lack an undrestanding of

[18] https://github.com/DataResponsibly/SynRD

the impacts of this technology. In summary, we are badly in need of broad educational initiatives that would be responsive to the learning needs of different stakeholders.

The *NYU R/AI* team has been developing educational materials and methodologies, and teaching responsible AI to members of the public [19], current and future data scientists [20], government accountability professionals, librarians, and practitioners at several large commercial entities. There is an urgent need to scale up these and similar educational efforts, to support a productive regime of distributed accountability.

Responsible data science and AI courses attract student cohorts that are engaged and demographically diverse. Stoyanovich and Lewis co-authored a paper describing course content, and its novel pedagogical methodology, in the International Journal of AI in Education [20]. For their course project, students work in teams to develop a nutritional label for an Automated Decision System (ADS) of their choice. These labels are intended as audits of these ADS, and can, for example, inform an organization's procurement team about an ADS they are about to deploy: does it work as advertised, is its training data appropriate, is it fit for use, does it have comparable accuracy across demographic groups, what might be some unintended effects of this ADS?

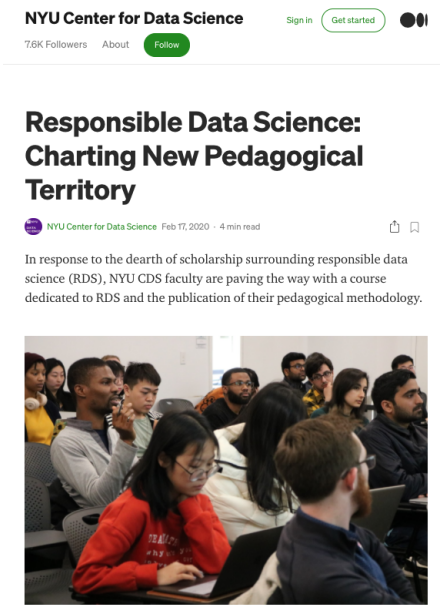


Figure 3: The Responsible Data Science course attracts a diverse and engaged cohort of students.

Materials and methodologies developed for the course at NYU have been useful to build additional materials for current data science professionals, notably, within city and state governments. The *NYU R/AI* team has delivered trainings to data scientists and decision makers in industry and government, including, for example, the Office of Innovation and Technology in the City of Philadelphia, the data science group at Fitch Ratings, and to the office of New York State Comptroller.

The *NYU R/AI* team is planning to co-create additional responsible data science and AI materials with organizations such as government and private sector entities, in ways that are targeted at specific roles within these organizations and based on concrete use cases, is in our immediate plans. Specifically, we see a need to develop technical hands-on trainings for data science professionals who develop, validate, and deploy AI solutions. We also see a need to train "mixed" teams of technical developers and less-technical product managers, enabling them to communicate about, and jointly address, ethical and legal challenges at the time when AI solutions are commissioned and built.

[19] https://dataresponsibly.github.io/we-are-ai/
[20] https://dataresponsibly.github.io/courses/

**25. Is the lack of a general federal data protection or privacy law a barrier to effective AI accountability?**

Yes, indeed. From the technical standpoint, it presents a barrier because having a federal data protection law would spur innovation in data minimization and in synthetic data generation, which would be helpful for accountability.

## AI Accountability Policies

**30. What role should government policy have, if any, in the AI accountability ecosystem?**

Governments — at every level — should serve as positive examples of AI accountability. They should teach their staff about the basics of responsible AI. They should establish, and then follow, AI governance frameworks. They should be proactive in assessing the systems they deploy (either built in-house or externally procured), and in making assessment results publicly available and accessible.

We also need strong laws, ideally at the federal level, to set standards for AI oversight, both in a sector-independent way and within each sector.

**31. What specific activities should government fund to advance a strong AI accountability ecosystem?**

It is crucial that the federal government funds responsible AI education initiatives. There are no alternative funding sources for this work, this work is badly needed, and it has to be done on the large scale. We cannot outsource this work to the industry, because of the conflict between societal and commercial incentives that simply cannot be overlooked in this case. US-based academic institutions are best-positioned to lead this work, in collaboration with others.

Further, it is crucial that the government support small and medium sized businesses in their transition into responsible AI. Education and upskilling, and legal and regulatory compliance are very expensive, and we risk placing smaller companies at an extreme commercial disadvantage.

# References


[1] Open provenance. https://openprovenance.org. [Online; accessed 14-August-2019].

[2] Serge Abiteboul, Gerome Miklau, Julia Stoyanovich, and Gerhard Weikum. Data, Responsibly (Dagstuhl Seminar 16291). Dagstuhl Reports, 6(7):42–71, 2016.

[3] Serge Abiteboul and Julia Stoyanovich. Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. ACM Journal of Data and Information Quality, 11(3):15:1–15:9, 2019.

[4] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. Fairness and friends. Data, Responsibly Comic Series, 2, 2021.

[5] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. Translational tutorial: Fairness and friends. In 4th Annual Conference on Fairness, Accountability, and Transparency, ACM FAccT, 2021.

[6] Falaah Arif Khan and Julia Stoyanovich. Mirror, mirror. Data, Responsibly Comic Series, 1, 2020.

[7] Abolfazl Asudeh, H. V. Jagadish, Gerome Miklau, and Julia Stoyanovich. On obtaining stable rankings. PVLDB, 12(3):237–250, 2018.

[8] Abolfazl Asudeh, H. V. Jagadish, and Julia Stoyanovich. Towards responsible data-driven decision making in score-based systems. IEEE Data Eng. Bull., 42(3):76–87, 2019.

[9] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In Proceedings of the 2019 International Conference on the Management of Data, SIGMOD, pages 1259–1276. ACM, 2019.

[10] Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Herasymova, Lucas Rosenblatt, and Julia Stoyanovich. The possibility of fairness: Revisiting the impossibility theorem in practice. In FAccT '23: 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA. ACM, 2023.

[11] Andrew Bell, Oded Nov, and Julia Stoyanovich. The algorithmic transparency playbook: A stakeholder-first approach to creating transparency for your organization's algorithms. In Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, and Anicia Peters, editors, Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA 2023, Hamburg, Germany, April 23-28, 2023, pages 554:1–554:4. ACM, 2023.

[12] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In Proceedings of the 5th Annual ACM Conference on Fairness, Accountability, and Transparency, FAccT, pages 248–266. ACM, 2022.

[13] Lucius Bynum, Joshua R. Loftus, and Julia Stoyanovich. Disaggregated interventions to reduce inequality. In EAAMO 2021: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Virtual Event, USA, October 5 - 9, 2021, pages 2:1–2:13. ACM, 2021.

[14] Marina Drosou, H. V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in Big Data: A review. Big Data, 5(2):73–84, 2017.

[15] Stefan Grafberger, Paul Groth, Julia Stoyanovich, and Sebastian Schelter. Data distribution debugging in machine learning pipelines. The VLDB Journal — The International Journal on Very Large Data Bases (Special Issue on Data Science for Responsible Data Management, 2021.

[16] Stefan Grafberger, Julia Stoyanovich, and Sebastian Schelter. Lightweight inspection of data preprocessing in native machine learning pipelines. In CIDR 2021, 11th Conference on Innovative Data Systems Research, Online Proceedings. www.cidrdb.org, 2021.

[17] Yifan Guan, Abolfazl Asudeh, Pranav Mayuram, H. V. Jagadish, Julia Stoyanovich, Gerome Miklau, and Gautam Das. MithraRanking: A system for responsible ranking design. In Proceedings of the 2019 International Conference on the Management of Data, SIGMOD, pages 1913–1916. ACM, 2019.

[18] H.V. Jagadish, Julia Stoyanovich, and Bill Howe. The many facets of data equity. ACM Journal of Data and Information Quality, 14:1–21, 2023.

[19] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. Towards substantive conceptions of algorithmic fairness: Normative guidance from equal opportunity doctrines. In Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022, pages 18:1–18:10. ACM, 2022.

[20] Armanda Lewis and Julia Stoyanovich. Teaching responsible data science. International Journal of Artificial Intelligence in Education (IJAIED), 2021.

[21] Anthony McCosker, Xiaofang Yao, Kath Albury, Alexia Maddox, Jane Farmer, and Julia Stoyanovich. Developing data capability with non-profit organisations using participatory methods. Big Data & Society, 9(1):20539517221099882, 2022.

[22] Vera Zaychik Moffitt, Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. Collaborative access control in WebdamLog. In Proceedings of the 2015 International Conference on the Management of Data, SIGMOD, pages 197–211. ACM, 2015.

[23] Luc Moreau, Bertram Ludäscher, Ilkay Altintas, Roger S. Barga, Shawn Bowers, Steven P. Callahan, George Chin Jr., Ben Clifford, Shirley Cohen, Sarah Cohen Boulakia, Susan B. Davidson, Ewa Deelman, Luciano A. Digiampietri, Ian T. Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole A. Goble, Jennifer Golbeck, Paul T. Groth, David A. Holland, Sheng Jiang, Jihie Kim, David Koop, Ales Krenek, Timothy M. McPhillips, Gaurang Mehta, Simon Miles, Dominic

Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Emanuele Santos, Carlos Eduardo Scheidegger, Karen Schuchardt, Margo I. Seltzer, Yogesh L. Simmhan, Cláudio T. Silva, Peter Slaughter, Eric G. Stephan, Robert Stevens, Daniele Turi, Huy T. Vo, Michael Wilde, Jun Zhao, and Yong Zhao. Special issue: The first provenance challenge. Concurrency and Computation: Practice and Experience, 20(5):409–418, 2008.

[24] Haoyue Ping, Julia Stoyanovich, and Bill Howe. DataSynthesizer: Privacy-preserving synthetic datasets. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM, pages 42:1–42:5. ACM, 2017.

[25] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. An external stability audit framework to test the validity of personality prediction in ai hiring. Data Mining and Knowledge Discovery, Special Issue on Bias and Fairness in AI", 2022. forthcoming.

[26] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich. Resume format, linkedin urls and other unexpected influences on ai personality prediction in hiring: Results of an audit. In Proceedings of the Fifth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AI 2022.

[27] Luke Rodriguez, Babak Salimi, Haoyue Ping, Julia Stoyanovich, and Bill Howe. MobilityMirror: Bias-adjusted transportation datasets. In Proceedings of the 1st Workshop on Big Social Data and Urban Computing, BiDU at VLDB, volume 926 of Communications in Computer and Information Science, pages 18–39. Springer, 2018.

[28] Lucas Rosenblatt, Joshua Allen, and Julia Stoyanovich. Spending privacy budget fairly and wisely. CoRR, abs/2204.12903, 2022.

[29] Lucas Rosenblatt, Anastasia Holovenko, Taras Rumezhak, Andrii Stadnik, Bernease Herman, Julia Stoyanovich, and Bill Howe. Epistemic parity: Reproducibility as an evaluation metric for differential privacy. CoRR, abs/2208.12700, 2022.

[30] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. FairPrep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In Proceedings of the 23nd International Conference on Extending Database Technology, EDBT, pages 395–398, 2020.

[31] Sebastian Schelter and Julia Stoyanovich. Taming technical bias in machine learning pipelines. IEEE Data Eng. Bull., 43(4), 2020.

[32] Mona Sloane, Ian Solano-Kamaiko, Jun Yuan, Aritra Dasgupta, and Julia Stoyanovich. Introducing contextual transparency for automated decision systems. Nature Machine Intelligence, 5:187—195, 2023.

[33] Julia Stoyanovich. Hiring and AI: Let job candidates know why they were rejected. The Wall Street Journal, 09 2021.

[34] Julia Stoyanovich, Serge Abiteboul, Bill Howe, H. V. Jagadish, and Sebastian Schelter. Responsible data management. Commun. ACM, 65(6):64–74, 2022.

[35] Julia Stoyanovich and Falaah Arif Khan. All about that bias. We are AI Comic Series, 4, 2021.

[36] Julia Stoyanovich and Falaah Arif Khan. Learning from data. We are AI Comic Series, 2, 2021.

[37] Julia Stoyanovich and Falaah Arif Khan. We are AI. We are AI Comic Series, 5, 2021.

[38] Julia Stoyanovich and Falaah Arif Khan. What is AI? We are AI Comic Series, 1, 2021.

[39] Julia Stoyanovich and Ellen P. Goodman. Revealing algorithmic rankers. Freedom to Tinker, Center for Information Technology Policy, Princeton University, 8 2016.

[40] Julia Stoyanovich and Bill Howe. Refining the concept of a nutritional label for data and models. Freedom to Tinker, Center for Information Technology Policy, Princeton University, 5 2018.

[41] Julia Stoyanovich and Bill Howe. Nutritional labels for data and models. IEEE Data Eng. Bull., 42(3):13–23, 2019.

[42] Julia Stoyanovich, Bill Howe, Serge Abiteboul, Gerome Miklau, Arnaud Sahuguet, and Gerhard Weikum. Fides: Towards a platform for responsible data science. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM, pages 26:1–26:6. ACM, 2017.

[43] Julia Stoyanovich, Bill Howe, and H.V. Jagadish. Responsible data management. PVLDB, 13(12):3474–3489, 2020.

[44] Julia Stoyanovich, Mona Sloane, and Falaah Arif Khan. Who lives, who dies, who decides? We are AI Comic Series, 3, 2021.

[45] Julia Stoyanovich, Jay J. Van Bavel, and Tessa V. West. The imperative of interpretable machines. Nature Machine Intelligence, 2:197–199, 2020.

[46] Julia Stoyanovich, Ke Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In Proceedings of the 21th International Conference on Extending Database Technology, EDBT, pages 241–252. OpenProceedings.org, 2018.

[47] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. MithraLabel: Flexible dataset nutritional labels for responsible data science. In Proceedings of the 28th International Conference on Information and Knowledge Management, CIKM, pages 2893–2896. ACM, 2019.

[48] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI, pages 6035–6042. ijcai.org, 2019.

[49] Ke Yang, Biao Huang, Julia Stoyanovich, and Sebastian Schelter. Fairness-aware instrumentation of preprocessing pipelines for machine learning. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA at SIGMOD, 2020.

[50] Ke Yang, Joshua Loftus, and Julia Stoyanovich. Causal intersectionality and fair ranking. In Symposium on the Foundations of Responsible Computing FORC, 2021.

[51] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM, pages 22:1–22:6. ACM, 2017.

[52] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. A nutritional label for rankings. In Proceedings of the 2018 International Conference on the Management of Data, SIGMOD. ACM, 2018.

[53] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, Part I: Score-based ranking. ACM Comput. Surv., apr 2022.

[54] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, Part II: Learning-to-rank and recommender systems. ACM Comput. Surv., apr 2022.